SRPeek: Super Resolution Enabled Screen Peeking

***Abstract*— We use smartphones everywhere and privacy concerns have arisen for the "shoulder surfing" attack, where strangers peek at our phone in public areas. To mitigate this privacy threat, various countermeasures have been designed but mainly against attackers with the naked eye. With the development of smartphones and super resolution (SR) technique, a malicious attacker can peek from further away with the assist of his/her smartphone camera and SR algorithms, rendering the underestimated threat model. To underline this concern, we present SRPeek, an end-to-end system of shoulder surfing deployed on commercial smartphones, including a unique deep neural network (DNN) architecture for multi-image SR. We implement SRPeek in Android. The experimental results demonstrate its efficiency and robustness, allowing a human to read 90% of the characters at a distance of 1.8m with traditional lenses and 6m with optical zooming lenses. And it outperforms the state-of-the-arts and fills the blank space of multi-image SR for shoulder surfing attack.**

I. Introduction

Smartphones have become a necessity in our lives, as we are checking our mobile phones constantly throughout the day. As a result, some privacy concerns have arisen about nearby parties peeking at our screen, namely "shoulder surfing". Given that the attacker is not malicious and merely takes several peeks out of curiosity, most works in this area are built on a threat model of an attacker observing with naked eyes, avoiding the severe data leakage effectively [6]. For example, a mere stranger can hardly do any harm reading a fragment of the correspondence or acquiring the password shown on screen without any equipment.

The privacy concerns however have been exacerbated with the rapidly developed mobile camera module these years. Equipped with multiple cameras, the newest generation of smartphones can perform 100× zooming compared to the standard 5× of single camera phones. Hardware improvements in memory also allows the burst mode at tremendous frame rates and even high-speed photography, delivering videos with thousands of frames per second. Given the commercial smartphone, the attacker can record the information reliably for propagating. And more images mean more recorded information of the victim. For example, in a Senate hearing, the Justice Secretary of Philippines, Vitaliano Aguirre II, suffered a leakage of his text messages, as someone had taken a snapshot of his smartphone [21]. To gain vital data more stealthily and accurately from even further away, the processing technique of multi-frame super resolution (SR) algorithms can be further employed by the attacker, delivering a long-range shoulder surfing threat model.

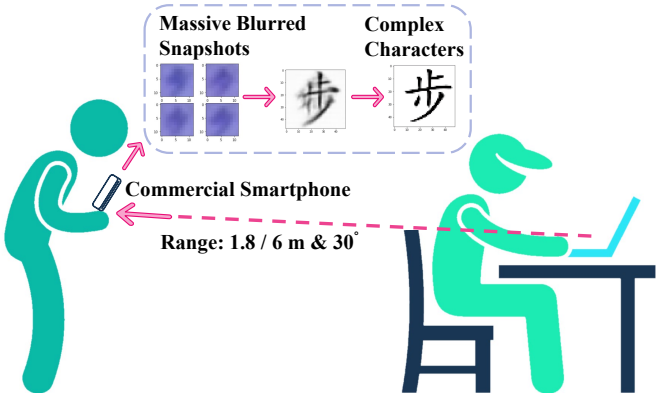


Fig. 1: Illustration of the long-range shoulder surfing threat model on smartphones, exacerbated by the SR algorithm.

Unfortunately, none of existing work can mitigate this new threat model. Specifically, massive efforts have been made, from physical privacy films to alternative password entry interfaces [20], [27] and input methods [13]. All of these methods however require the additional deployment cost [3] and cannot be widely deployed for critical privacy stages (e.g., password entry) of most scenarios. Furthermore, none of existing works can deal with the presence of developed smartphone cameras and SR algorithms in shoulder surfing scenarios. Given the ubiquitous usages of smartphones and serious damages of the data disclosure for the victim, it is imminent to recall attention on this new threat model and propose corresponding countermeasure for modern circumstances.

Threat Model. We present SRPeek, an end-to-end system of shoulder surfing using the SR algorithm, illustrated in Figure 5. While the victim's reading private information in public, the attacker, standing within 1.8 meters behind him, raises his commercial smartphone and zooms in on the victim's screen. Then he can take 10 to 20 snapshots rapidly in burst mode (about 10 frames per second), and process them with a multi-frame SR technique to generate a high-resolution result. The whole process only takes a few seconds and can even be repeated at a 2 second interval, enabling the attacker to monitor the victim's screen continuously, not missing any detail. Meanwhile, few people would suspect a stranger holding his smartphone at 1.8 meters behind them. Given the newest generation of smartphones equipped with telephoto lenses and capable of optical zooming, this distance can reach a staggering 6 meters, posing a silent but deadly threat to screen privacy.

Challenges. Creating a powerful SR based shoulder surfing threat model, however, is not trivial, especially for commercial smartphones in real time. And we achieve SRPeek by overcoming four key challenges.

- **Blurriness of input images.** The most prominent one is the blurriness of captured snapshots, as they are taken secretly and hurriedly at extreme range, without specially focusing. It can be further exacerbated by straining the magnification of the smartphone lenses. Unlike most SR applications and datasets working with "normally" captured snapshots (e.g., scenery, scanned images) [17], [19], these snapshots exhibit lower concentrations of information and more artifacts as Figure 5 shows, requiring a reliable "reconstruction" than "interpolation". Since blurrier input means less information, rendering a greater possibility of reconstructing the wrong character. On one hand, each snapshot is blurred with a randomly different PSF kernel, removing the consistency between neighboring frames. Thus, it cannot be approximated as video clips using a constant, isotropic gaussian kernel; on the other hand, each defocused snapshot only contains fractions of information, analogous to pieces of a jigsaw puzzle, posing a greater demand for integration abilities of the SR algorithms than most other SR applications.
- **Super resolution with multiple snapshots.** To resolve the blurriness of input snapshots, a SR algorithm on a smartphone is required to recover the blurred contents. However, none of existing SR algorithms can be adapted for this attack scenario, especially with massive extremely blurred snapshots as input. Since smartphones can capture 10 snapshots per second in burst mode. Specifically, the difference between videos and multiple snapshots rules out existing image/video super resolution algorithms [11], [16]. The reason has two folds. First, massive input leads to the increase in model complexity, making it unsuitable for mobile deployment. Second, the nature of our task requires to do pixel-wise correlation and comparison in the same coordinate with consecutive snapshots throughout our construction process. Otherwise we cannot tell if, for example, it is one stroke or two parallel strokes that is shown in these two darker pixels on this snapshot.
- **Model complexity and variable input.** We have to operate our threat model locally on commercial smartphones in a real-time manner rather than remotely in cloud server. Since 1) the latter requires a stable internet access, and 2) sending multiple snapshots across the Internet will consume time and bandwidth. Given the required 10 to 20 snapshots to be processed for the ideal content recovery, we have 1-2 seconds as the minimum interval between each output frame, rendering a limited period for calculation with the restricted power and RAM. In case where the display on the target screen changes more frequently, fewer snapshots and less processing time are available for each scene. Consequently, it cannot only require a more precise network model, but also a model that can deal with dynamic inputs efficiently, either 3 or 20 snapshots.
- **Complexities of characters.** To deploy our threat model broadly for real-life scenarios (e.g., record passwords or e-mails), it is supposed to reconstruct complex characters with multiple strokes, such as Chinese characters shown in Figure 5. Composed of multiple strokes, characters are distributed discretely in the pixel-wise snapshot and different with other entities (e.g., faces and natural objects), rendering unique challenges. In some cases, messing up a stroke even slightly can shorten or lengthen a stroke, leading to a different character for misunderstandings. Unlike most SR applications working on similarity between their output and the ground truth, our network architecture and training process must be engineered to recover readable contents. Another challenge is the imbalanced element for words. Since we know certain stroke or alphabet of Chinese and English characters can occur more frequently, results in lower prediction accuracy over the less common, unconventionality shaped characters.

Summary of Evaluation Results. Equipped with our uniquely designed SR algorithm, our system can help the attacker read the characters shown on the victim's screen with above 90% accuracy at 1.8m distance on a smartphone without optical zooming, or 6m on a phone with optical zooming in our experimental settings. SRPeek can capture snapshots and produce high-resolution images constantly with about 2 seconds interval. Our experiments in real life scenarios show that it can decipher texts or passwords of the victim at a safe distance, without alarming the victim, posing a serious threat to screen privacy.

Contributions. This paper makes the following contributions:

- We propose SRPeek, an end-to-end threat model of shoulder surfing from a broad range, which can be deployed on commercial smartphones. To the best of our knowledge, we are the first to consider the presence of smartphone cameras and SR algorithms in shoulder surfing scenarios.
- We design a multi-frame SR neural network architecture to reconstruct massive extremely blurred and defocused snapshots. It can be further extended to other text-recognition applications.
- We evaluate this new shoulder surfing threat model in multiple scenarios. And it outperforms the state-of-the-arts for content recognition, calling for the new privacy concern.

The rest of the paper is organized as follows. Section II describes the related work, especially the state-of-the-arts for the shoulder surfing and SR techniques. We present the system and network design in Section III and IV, followed by the implementation in Section V and evaluation in Section VI and

TABLE I: A comparison of the state-of-the-arts on shoulder surfing.

Reference	Scenarios	Metric	Quantitative Model	Distance[a]
Eiband [6]	Naked eye	×	×	1m
Kwon [14]	Naked eye	×	×	-[b]
Schaub [24]	Naked eye	✓	×	-
Maggi [18]	Camera	✓	×	-
SRPeek	COTS phones	✓	✓	1.8 / 6m

[a]The maximum distance to the victim's screen. [b]The attacker stands next to the victim without the maximum effective distance.

VII. We further wrap up this paper by delivering the limitations and countermeasures of our system in Section VIII. And the conclusion is shown in Section IX.

II. Related Work

A. Shoulder Surfing

With the arrival of the information era, privacy issues are becoming increasingly prominent. Smartphone screen privacy, the concern of our smartphones being observed by strangers in public areas, or shoulder surfing, have been studied heavily recently [6], [9], [14]. To mitigate this threat, some systems hide the information [1] or warn the user [22] once sensing malicious passers-by; others modify the user interfaces, including creating honeypots (for passwords) [2], confusing unauthorized parties [27], and making the interactions invisible [13] or unreadable from a distance [3]. Most of the works assume that the attacker is a casual passer-by, taking occasional peeks with the naked eye, as is the case most of the time [6] [27]. Given the assisted equipment, the malicious attacker can however acquire the sensitive information (passwords, business correspondence, etc.) readily to do real harm. To deal with the threat model of tool-assisted shoulder surfing,Maggi et al. designed an automatic shoulder surfing threat model, observing the target smartphone with a camera [18]. It can however only function when the attacker is standing at close range, which is a barely practical scenario. By tailoring and fusing the SR technology with smartphones with developed mobile camera module and processing ability, we propose a stronger shoulder surfing threat model in which attackers can deploy it on commercial smartphones while obtaining information for a longer range to reduce suspicion, say 1.8-6m away from the victim's screen with an observing view of 30° shown in Table I. Evaluations demonstrate its feasibility and privacy concerns in our daily life.

B. Super Resolution

Image SR is the process of reconstructing an image with a higher spatial resolution. Based on the structural pattern, self-similarity [26], or previous knowledge of the image genre, the single-image SR techniques take as input a single low-resolution image, rendering a sharp, high-resolution one by deducing missing information and reconstructing the missing pixels. Further, multi-image SR techniques work on a set of pictures on the same scene, such as multiple snapshots captured by a smartphone, successive images from a satellite, or adjacent frames on a video clip. These algorithms collect extra data from slight differences of these redundant images, often exhibiting better performance than single-image SR algorithms. And a variety of techniques have been proposed for the multi-image SR problem. And early works focus on the analysis of images in spatial or frequency domains. For example, the Shift-Add algorithm [7] analyzes the spatial information by maximizing the pixel-wise possibility of high-resolution image from low-resolution images. Besides, approaches for the frequency domain rely on the Fourier transform of images. Assuming that the high-resolution scene is band-limited, these algorithms reconstruct high-frequency components from patterns of low-frequency components, removing bands of noise and blur effects simultaneously. Unfortunately, all SR techniques however, are subject-specific and achieve the best performance only for some contents, including natural photos, scanned text, satellite imaging, biometrics [28] . While SRPeek can recover multi-type characters with complex strokes, including Chinese, English and numbers.

Deep learning approaches can also be adopted for Super Resolution. For example, SRCNN [4] was designed using CNN by replacing pooling layers with upsampling layers, achieving notably better performance than previous approaches. It however only focuses on single image SR, as the neural network accepts a single image as input. Besides, Generative Adversarial Networks (GAN) were also used to generate photo-realistic images [15], with less artifacts and more genuine-looking details perceptible to the human eye. To resolve multi-image SR tasks, say video SR, existing works [11], [25] design a structure similar to SRCNN, and accept complete video clips as input. Then the sequentially and consistency between adjacent frames can help recover the missing pixels of clips using convolution networks, Specifically, we can modify the dataflow to merge neighboring frames among the network layers [10], or recurrently processing the frames under the guidance of the output of the previous frame [23]. For images without consistency or sequential information, like satellite images, most works choose hybrid methods to solve the multi-image SR problem with multiple single-image SR procedures. They either merge the results of single-image SR algorithms for efficiency [12], or build a multi-image network to create a comprehensive view based on single image SR networks [5]. Due of the extreme blurriness of snapshots, these methods cannot deal with the new shoulder surfing threat model we proposed, which takes as input massive blurred snapshots without the consistency between frames. And they achieve limited performance, shown in Table I.

III. System Overview

We implemented a holistic system for shoulder surfing, with the neural network as core, on a smartphone to verify the

efficiency of our model. It iterates through the following steps:

Input: Under guidance of the attacker, The smartphone will zoom in, take focus, and take 20 images with burst mode of the target screen. Note that zooming is only for easier interaction and focusing, and to utilize the telephoto lenses and optical zoom, if available, as digital zooming do not put in any extra data; On traditional phones with only digital zooming, the photo-taking will be with 1x zoom, and on phones with optical zooming, 5x to 10x zoom, depending on the optical zooming range. This way information on the images will be more compact and easier to comprehend by neural networks.

Alignment: The images are then aligned to mitigate the shifts between frames caused by hand tremor or movements from the lenses, as cameras with optical lenses will occasionally shift slightly across time due to the movements of its inner mechanism. Luckily, in our scenario the target is a glowing screen whose edges are easily distinguishable in most cases, and we use them as reference to align our images. The images are also spun to make the text horizontal in the process. The screen is cropped out and the rest of the image abandoned.

Adjustment: The lines of text(or stuff differing from the background color of the screen) will be carved out for processing, to reduce the workload of the network. The characters are normally around 10x10 pixels in the photos, and the carved segments will leave 2 pixels of padding on all sides to avoid mutilating the character; A certain amount of error is allowed, but if the size of the text is too small or too large, the neural network will not be able to extract features normally, so the images will be zoomed to the right size (and the zooming of the camera lens in the Input phase will be adjusted accordingly).

Processing: As our network accepts only 9x9 patches, we will process all the 9x9 patches among the input photo with our multi-frame super resolution network to generate 36x36 images (4x upscaling). The input is RGB colored, while the output, as we are not interested in color, is black and white. In this way, we generate upscaled image segments of all the characters on the target screen. The three calculation processes (Alignment, Adjustment, Processing) runs parallel to the input process, so that once a batch of input images is available, they can be calculated and displayed on screen with a slight lag while the attacker can start to collect the next batch at the same time.

Output: The outputs of our SR network, the 36x36 patches, are then merged into a whole. The overlapping pixels are thus processed: among all the outputs containing this pixel, we collect these values, remove outliers, and use their average as the result. As we only carve out the texts on the phone and process them, we will then insert these high-res texts back to their original locations on one of the input images, and display the patched image to the attacker (An illustration can be found in Fig. 5).

These steps are repeatedly executed to enable the attacker to monitor the victim at an interval of a few seconds. At critical times requiring continuous monitoring so as not to miss transient display, e.g. password entering, the system can simply lengthen the input phase across that period and process the data afterwards. The workflow of our system is shown in Figure 2.

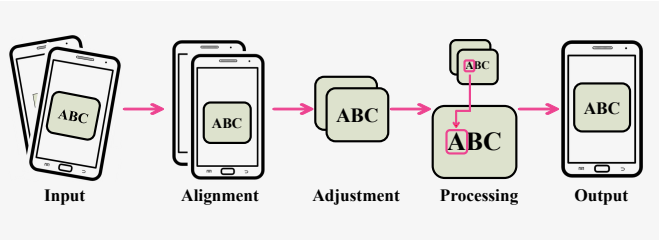


Fig. 2: Illustration of the workflow of SRPeek.

IV. System Design

A. Design Principal

To solve the challenges and outperform the state-of-the-arts for this new shoulder surfing attack, we propose a holistic system with the multi-frame SR neural network, illustrated in Figure 2.

- **Layered architecture and frequent introduction of input.** To counter the blurriness of the images, and the tendency of reconstructing fake characters, our network is constructed with several identical SR layers, connected sequentially, and the images are refined step by step throughout the layers. All input images are processed simultaneously and separately, with the information from other images for reference, throughout the network. Another important feature of our architecture is that the initial input images are introduced into the dataflow at each layer, from beginning to end, resulting in the uneven depth of our model, acting as an anchor for the reconstruction process so that the output will be faithful to truth from beginning to end, while preserving the deep mainstream of the model to be able to process such blurry images.
- **Merging layers to adapt multiple frames.** A merging process inside each layer functions as the revenue of communication across images integrates information from all the images into one, the result of which is then stacked with each image for the next convolution. At each layer the image is exposed to the consensuses of features from other images at the same level of complexity, acting as a verification for the hypothetical features extracted from the current image, so that in the training process more audacious features can be learned and proposed without fear of punishment from the final loss, increasing the quality of featuremaps throughout the network.
- **Model Complexity and adaption for variable input.** All the convolutional layers throughout the network

are performed individually for each image (or the featuremaps of them from the previous layer), and these convolutions in the same layer share a same set of parameters in training and production, thus reducing parameter count and avoiding convolutions for large numbers of channels, saving calculation power. This trait also enables the same network to adapt any number of input images. On one hand, this gives our model the ability to adapt to the uncertainty of the number of available images, providing relatively satisfactory results regardless whether the attacker has ample time photographing the target phone before it changes its display; on the other hand, the relative independence and the isotropy of the merging layers avoids the reliance on sequential order and consistency between neighboring frames(which is common among most multi-frame SR networks), solving the inconsistency problem mentioned in Figure 5.

- **Compatibility with complex characters.** The discrete distribution of characters leads to the inclination of deviation and producing 'fake' results. The frequent introduction of input images is designed to mitigate this challenge. Also, our model can learn correlations between certain features and strokes from the training data, decomposing the SR problem and making it a lot less challenging. We also introduce adaptive boosting [8] in our training process, increasing the loss weight of the wrongly reconstructed characters, determined by loss in earlier stages and OCR in later stages, to accommodate the imbalances of data. The loss functions in training is also redesigned. The mean square error (MSE) is not suitable in this application, as a misplaced stroke, although highly obstructing readability, may only invoke a slight decrease in MSE because it only influenced a few pixels. Our solution is putting a weight on each pixel before applying weighted MSE. Assuming the characters are black on white, this weight increases at darker pixels and propagates to neighboring dark pixels so that long strokes and intersections of strokes are given higher weights. This process serves as a supplement to MSE to focus more on readability and is beneficial for OCR and human reading tests.

B. Network Design

The core of the SRPeek system is a specially designed multi-frame SR neural network, accepting a group of N images indexed $x_1^{(0)}$ to $x_N^{(0)}$ as input and generating an image with higher resolution y as output. The network comprises L layers, each of which implements a non-linear transformation $H_l(\cdot)$, where l indexes the layer. As mentioned before, in each layer images are processed separately, with the merging layers as a revenue for communication, so that the output of each layer is correspondent to the input. We denote the output of the l^{th} layer as $x_1^{(l)}$ to $x_N^{(l)}$, which is also the input of the $(l+1)^{th}$

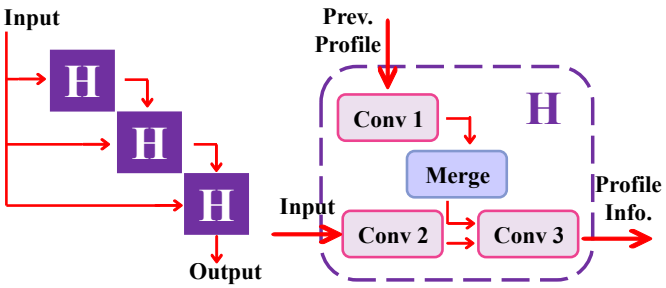


Fig. 3: Network Architecture of SRPeek.

layer. Till now the model is not different from traditional SR models:

$$x_i^{(l)} = H_l(x_i^{(l-1)}), i = 1, 2, ...N \tag{1}$$

Additionally, we introduce the initial images $x_i^{(0)}$ as an input for all the layers:

$$x_i^{(l)} = H_l(x_i^{(l-1)}, x_i^{(0)}), i = 1, 2, ...N \tag{2}$$

The last layer is exceptional, it yields a single image y as output.

Presently, in these layers nothing is done to raise the resolution of the images, so that the resolution of $x_i^{(0)}$ to $x_i^{(l-1)}$ and y remains the same. To increase resolution, we insert several $2\times$ nearest upsampling layers U evenly throughout the architecture between the layers:

$$x_i^{(l)} \leftarrow U(x_i^{(l)}), x_i^{(0)} \leftarrow U(x_i^{(0)}), i = 1, 2, ...N \tag{3}$$

we upsample the input images $x_i^{(0)}$ simultaneously to keep the two inputs of the following layers $H_l(x_i^{(l-1)}, x_i^{(0)})$ unanimous in resolution. For example, in a $4\times$ SR network with five layers, we may insert 2 $2\times$ nearest upsampling layers behind the 2nd and 4th layer.

Inside each layer H_l there are 3 convolution layers $Conv_{1l}(\cdot)$, $Conv_{2l}(\cdot)$, $Conv_{3l}(\cdot)$ and 1 merging layer $Merge_l(\cdot)$.

Conv1: The first convolutional layer, $Conv1$, accepts the layer's first input parameter $x_i^{(l-1)}$ as input. Note that all three convolutional layers accept a single image (or its featuremaps from the last convolutional layer) as input, the convolutional process is repeated for all the images, and calculations within the same convolutional layer share the same group of parameters all the time (parameters denoted as $Param_{sl}$ for convolutional layer $Conv_{sl}, s = 1, 2, 3$):

$$a_i^{(l)} = Conv_{1l}(x_i^{(l-1)}, Param_{1l}), i = 1, 2, ...N \tag{4}$$

Merge: The results of the previous step of all the images $\{a_1^{(l)}, a_2^{(l)}, ..., a_n^{(l)}\}$ are then passed to the merging layer $Merge_l$ to generate t groups of featuremaps. Suppose the results of $Conv_{1l}$ consists of R channels:

$$a_i^{(l)} = \{a_{i1}^{(l)}, a_{i2}^{(l)}, ..., a_{iR}^{(l)}\}, i = 1, 2, ...N \tag{5}$$

The data in each channel will be merged separately in the merging layer. The output is $T \times R$ channels, denoted as $b_{tr}^{(l)}(t = 1, 2, ...T, r = 1, 2, ...R)$:

$$b_{tr(p,q)}^{(l)} = \sum_{i=1}^{N} a_{ir(p,q)}^{(l)} e^{k_t a_{ir(p,q)}^{(l)}} / \sum e^{k_t a_{ir(p,q)}^{(l)}} \tag{6}$$

where (p,q) represent the pixel at this coordinate, and k_t is a set of fixed parameters shared in all the merging layers throughout the model, controlling the behavior of the merging process. Apparently, k=0 leads to averaging, k=$+\infty$ leads to max operator and k=$-\infty$ leads to min operator. We use T=5 and k=-1,-0.5,0,0.5,1 in our model, giving consideration to both consensuses (k=0,averaging) and prominent features(k=1, 'soft'max and k=-1, 'soft'min). these $T \times R$ channels $b_{tr}^{(l)}$ is the output of this merging layer $Merge_l$.

Conv2: $Conv2$ is a replica of $Conv1$, processing the layer's second input parameter $x_i^{(0)}$, also generating N outputs with R channels per output, denoted as $c_i r^{(l)}, i = 1, 2, ...N, r = 1, 2, ...R$:

$$\begin{aligned} c_i^{(l)} &= Conv_{2l}(x_i^{(0)}, Param_{2l}), i = 1, 2, ...N \\ c_i^{(l)} &= \{c_{i1}^{(l)}, c_{i2}^{(l)}, ..., c_{iR}^{(l)}\}, i = 1, 2, ...N \end{aligned} \tag{7}$$

Conv3: The data from $Merge$ and $Conv2$ are merged together, in that all the $T \times R$ channels of $b_{tr}^{(l)}$ are replicated N times and stacked with each one of the N outputs of $Conv_{2l}$, before these N outputs, each with $(T + 1) \times R$ channels, are passed through the third convolutional layer $Conv_{3l}$. There are also N output of this convolutional layer, denoted as $d_i^{(l)}, i = 1, 2, ...N$:

$$d_i^{(l)} = Conv_{3l}(Stack(c_i^{(l)}, b^{(l)}), Param_{3l}), i = 1, 2, ...N \tag{8}$$

Output: If $l < L$, this is not the last layer, the N outputs of step (4) will be the output of layer H_l. Otherwise, as we need to present a single image y as output, we add another merging and a common convolutional layer after $Conv_{3l}$. The merging layer is identical to the previous $Merge_l$, merging the N outputs $d_i^{(l)}$ into $T \times R$ channels $e_{tr}^{(l)}$, and a convolutional layer processes this $e_{tr}^{(l)}$ to generate a single channel of output y.

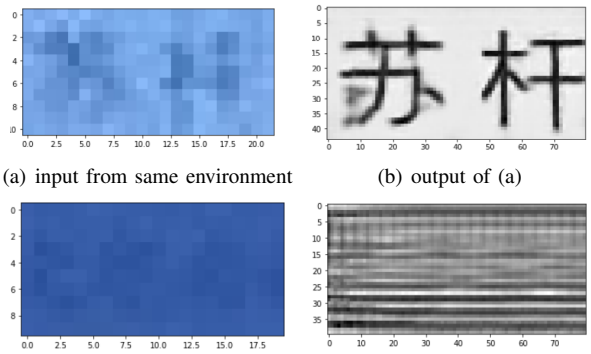
$$\begin{aligned} \{e_{tr}^{(L)}, t \le T, r \le R\} &= Merge'_L(\{d_i^{(L)}, i \le N\}) \\ y &= Conv'_L(\{e_{tr}^{(L)}, t \le T, r \le R\}) \end{aligned} \tag{9}$$

The full structure of a 3-layer network is shown in Figure 3.

V. Implementation & Training

A. Data Collection

In our experiment, we are forced to collect the training dataset on our own instead of using public datasets. Because of our unique application, to the best of our knowledge there is no publicly available image dataset built for shoulder-surfing.



(a) input from same environment (b) output of (a)

(c) input from different environment (d) output of (c)

Fig. 4: Results of training with data collected in fixed environment. The network performs well in this environment but fails in other.

Modern network architectures work with naturally obtained images, e.g. ordinary photos, satellite images, scanned documents, YouTube videos, etc. where the images are well focused and captured within the imaging ability of the camera, with the objects of interest, e.g. the face or the printed word, at least visible to the naked eye. and we apply the SR algorithms to reduce noise and refine details, like repainting texture and refining edges. However, as emphasized above, in our application we face images that is extremely blurred and distorted, due to the extreme circumstances when the photos are taken, and it is impossible to tell from each single picture whether a stroke of a character really belongs here, as it's equally possible that this line is a distorted mixture of multiple lines or is a mutilated part of a longer stroke. These are the reasons why our work does not choose datasets that are publicly available, and collect data by ourselves–as it will completely defeat the purpose if we turn to these datasets. And this also explains why we failed in trying to get comparable results with other SR architectures we use as baselines, as these networks are designed for another purpose and they hardly ever faced such distorted and blurred images.

The task of data collection is quite tedious, as we discovered that we need photos taken in all kinds of environments to train a robust network. We initially collect 60,000 images with the position of the smartphones static and all other environmental parameters stationary, and divided it into training and testing datasets which do not have common characters between them. Note that different lenses have different photographing abilities and blurring patterns, so the model has to be trained for each phone model. When using smartphones with optical zoom, the images will shift slightly time after time even if we keep them completely still, so the aligning phases are still needed (this shifting is also beneficial in that it simulates normal hand tremors in real-life scenarios even if we fix the phone to a stand instead of holding it up throughout the data collection phase which can last several hours). There is no consistency between

adjacent images, so we also shuffle the images to increase this shifting effect and increase the variety of the training data.

Our network easily learned the patterns and presented equally satisfactory results in the test dataset (see Figure 4(b)), displaying clean, easy to read results, showing that the network has learned to extract features beneath the character level, and it is highly possible that this network can recover all kinds of characters, not limited to the characters in the training dataset. However, when we slightly modified the environment, the images displaying the text with a different shade and size, none of the features were successfully extracted, and the network outputted white noise (see Figure 4(d)).

B. Experiment Setting

As a result of these observations, we have to collect images in varied environments. Our experiment consists of 2 smartphones, one for the attacker and one the victim. Their distance is between 1 and 2 meters for traditional lenses (for the camera of the attacker's phone) without optical zooming, and 5 to 7.5 meters for lenses with optical zooming. Both phones are fixed to stands to keep them completely still, but as mentioned above, the lenses with optical zoom will shift slightly time to time, which is similar to a handheld situation. An app runs inside the attacker's phone, taking photos continuously, adjusting its focus and aperture, trying it's best to capture high-quality photos. Another app runs inside the victim's phone, displaying random characters with several fonts and colors, for the attacker to capture. The characters are selected from commonly used Chinese characters with 5 to 10 strokes and the English alphabet, and we divide this assemble into training and testing subsets. As the English characters are apparently easier to classify and reconstruct for the networks, the experiments testing the performance of the network will be performed only on Chinese characters, but it is our observation that our system works on English characters as well as Chinese characters.

As the experiment goes on, the attacker phone will keep on taking photos in burst mode (20 photos per time), and the victim phone will change the characters it is displaying at a fixed interval, when approximately 100 images were taken by the attacker. We change the environmental setting whenever about 2k photos were taken, relocating the phones (while keeping their distance between 1 and 2 meters) or modifying the angle of the phones. The attacker will always point its phone directly at the victim, keeping its screen in the middle of the photos; The victim, however, may tilt its phone at an angle within 30 degrees. We then readjust the focus and aperture of the attacker's phone to maximize the image quality before continuing data collection. Screenshots were taken in the victim's phone each time it changes its display, and these screenshots are scaled, spun and deformed according to the distance and angle between the phones at the time, and used as ground truth for training and evaluation. An illustration of our experimental setting is displayed in Fig. 5.

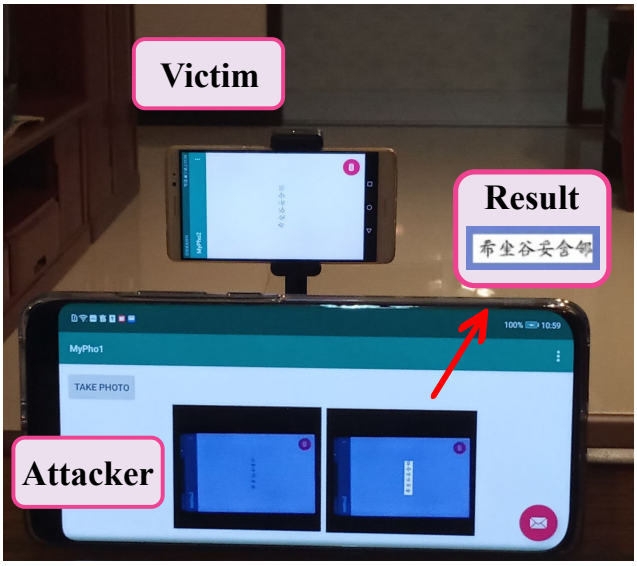


Fig. 5: Illustration of our experimental setting (Distance between phones is shortened for demonstration).

The data was collected indoors, as is often the case of shoulder-surfing in public areas (theaters, subways, offices, etc.). We perform the experiment in a room with a window and repeat the data collection phase in different times of the day and night, with curtains drawn or open, lights turned on and off to modify the illumination parameter. The position of the phones is also a crucial factor in this parameter, and we avoid extreme settings, e.g. having the sunlight directly shining on the screen. We collected 800,000 images in this way with a Redmi6A phone (which possess a camera with 130 million pixels). This process is extremely time-consuming, but the data amount is crucial in our experiment, as slight variations of the environment will cause drastic changes in the features extracted from the characters, and only by covering the variations in each of the environmental parameters in the training data can the system successfully function in different scenarios.

C. Model Specifications

It is our observation that cameras in different phones display different patterns of distortion when performing the shoulder-surfing experiment, and apparently, images captured from a phone with weaker abilities (less pixels, less range of focus, worn lenses, etc.) will need a stronger, more complex model to extract the features. Thus, the specifications of our model are only for reference and may not work when reproduced on another phone.

Our model accepts 20 images at a time, with a size of 9x9 pixels. The model consists of 5 blocks described in the previous section. In each block, feature maps from the previous layer are passed through a single convolutional layer with 32 channels of feature maps as output. the 20x32 feature maps are then merged horizontally with the max-min-average

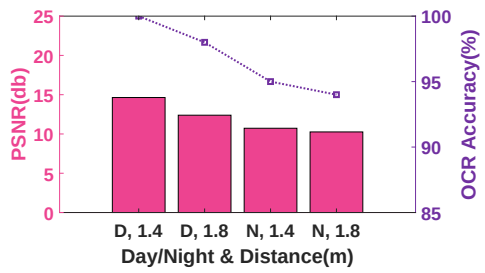
process, and the output is 5x32 channels. Simultaneously, the original input images are processed again with a single convolutional layer with 32 channels output, and stacked with the former 5x32 channels to form a dataflow of 6x32 channels for each one of the 20 images. These channels are finally passed through a single convolutional layer, outputting 32 channels per image for the next block. The kernels in each convolutional layer is 3x3. We also insert LeakyReLU and Batch Normalization processes after each convolutional layer, and 2 upsampling layers between the 5 blocks. A single 1x1 convolutional layer is placed after the 5 blocks with a single channel as output to form the final output layer. This model consists of approximately 200,000 parameters and a complexity of about 400,000 FLOPs. As a light-weighted model, it makes a prediction within 0.1 seconds on a Tesla K80 GPU over a 9x9 patch, and when implemented on a smartphone, the human user can recognize a character within 2 seconds of processing time.

D. Training Process

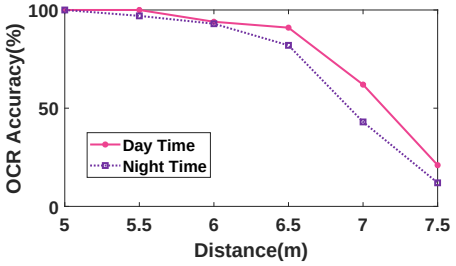
Because of the difficulties in discovering the patterns among the blurry images, the training process is not so straightforward and somewhat time-consuming. Our approach is as follows: we first collect 60,000 images at a stationary experiment setting and train the model with it, as mentioned in Sec V-A, until we get satisfactory results on the test dataset. The model should be able to handle this task easily. After that, we use transfer learning methods to fine tune the model in order to fit different environmental parameters. We add a small percentage of image data from another similar experiment setting into the training data and resume training. When the model stabilizes, continue adding data from the same setting until the model can equally process data from the two image collections. Note that in this process the model will tend to extract false features, displaying wrong but clear texts as result. This phenomenon can be mitigated with dropout and normalization layers. As the model successfully fits two of the image sets, we repeat this procedure for several times until it shows signs of self-adapting, for example, when presented photos taken at 1.8m and 1.2m range in the training process, the model fits 1.5m photos easily. After that, the model can learn with the full fully-shuffled dataset with fewer difficulties.

VI. Model Evaluation

We evaluate the performance of the network model in different training and testing conditions. We perform the following experiments with two COTS smartphones: a Redmi 6A smartphone, with a single rear camera with 13 million pixels and digital zoom only, and a HUAWEI P40 Pro, with multiple rear cameras. The telephoto camera possess up to 5x optical zooming ability which we will utilize fully in our experiments.



(a) Traditional Lens



(b) Optical Lens

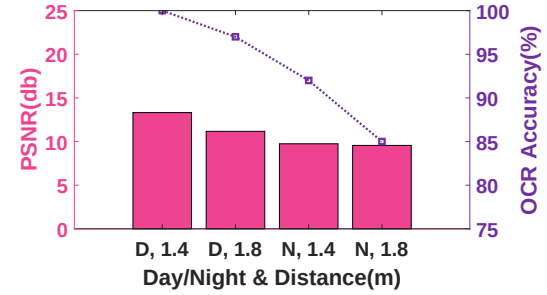
Fig. 6: Performance in Controlled Environment using Traditional Lens and Optical Lens.

A. Performance In Controlled Environments

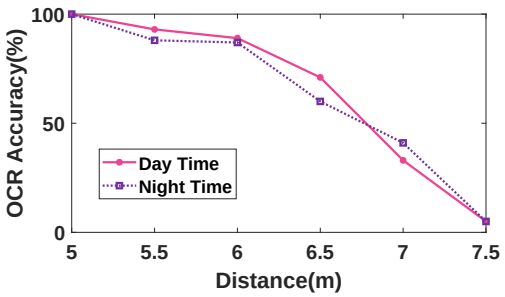
In these experiments, we train and test the model with the images captured with exactly the same environment parameters, and used Peak Signal to Noise Ratio to evaluate the accuracy of the recovered images. The results are shown in Figure 6(a) and Figure 6(b). Moreover, the ultimate goal of our system is the readability of the recovered images, so we also used Optical Character Recognition (OCR) services to evaluate the accuracy. The model with traditional lens without optical zoom (hereafter referred as traditional lens) is trained and tested at 1-2 meters range, while the lens with optical zoom (hereafter referred as optical lens) at 5-7.5 meters, at which distance less than 5% of the characters (only the simplest ones) can be recognized by humans from the photos without the assistance of SR algorithms. The PSNR and OCR results are shown in Fig. 6(a).

This model can achieve an OCR accuracy above 90% at 1.8m with traditional lens in both day and night time. The PSNR is higher than 10 as well. OCR accuracy and PSNR drops in night time since the darker environmental lighting condition as well as with the increasing of distance since the longer distance means more blurs. For Figure 6(b), we can see the OCR accuracy is larger than 90% at 6m with optical lens. The performance is relatively consistent in both day and night time. Considering the complexity of Chinese characters, and the assist of context when read by an OCR recognizer, we believe this level of accuracy can provide sufficient data at a shoulder surfing scenario, thus proving the efficiency of our model.

In the optical zoom group, the accuracy of the model

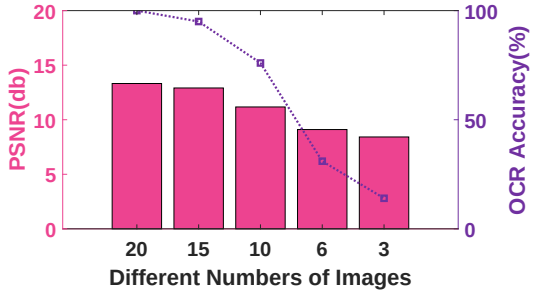


(a) Traditional Lens

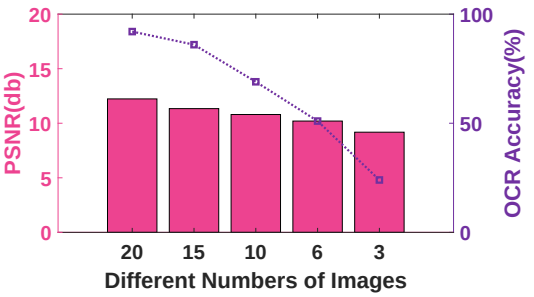


(b) Optical Lens

Fig. 7: Performance in Random Environment using Traditional Lens and Optical Lens.



(a) Traditional Lens



(b) Optical Lens

Fig. 8: Performance with fewer available images using Traditional Lens and Optical Lens.

dropped drastically at 7m distance. Increased distances mean less data and less restrictions of the possible outputs, which leads to artifacts (missing or misplaced strokes, etc.). It is the nature of Chinese characters that one mistaken stroke will largely affect its readability, leading to the result that while the pixel-wise error rises steadily with the increased distance, the accuracy will experience a drastic drop.

B. Performance In Random Environments

We train the model with data captured in different environments, as mentioned before, and test its ability in other environments. The results are shown in Figure 7(a) and Figure 7(b). For Figure 7(a), the model can still achieve an OCR accuracy above 85% at 1.8m with traditional lens. From Figure 7(b), the OCR accuracy keeps higher than 90% at 6m with optical lens. This verifies the efficiency of our model for environment adaption.

We compare the results to discover the influence of the surrounding environment: The model performs best in daytime and weaker illumination may decrease its ability. Also, the model performs better at closer distances. As the model is trained initially on the 5m group data, it performs especially well at this distance, while the accuracy at other distances experience certain levels of decrease. Also, similar to the experiments at controlled environments, random artifacts start to encompass OCR compensation, rendering the result images highly unreadable, until at 7.5m distance only the simplest characters can escape from being flooded by noise.

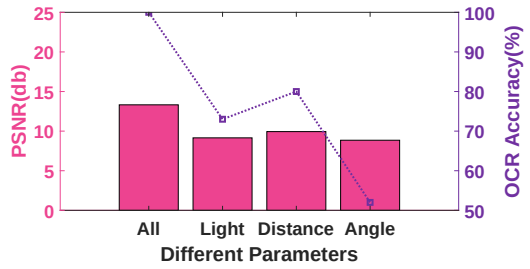
C. Performance with fewer available images

As mentioned in sec. IV, our model is designed to work on any number of input images, the only limitation being that these images are snapshots of the same scene. This design is requisite because in certain scenarios the data displayed on the victim's screen is transient and ever-changing, e.g. password entry, where only the last character of the password is visible. We evaluated the impact of fewer available images to the performance of the SR model. The results are shown in fig. 8. We tested at the 1.8m daytime scenario for traditional lens and 6m daytime scenario for optical lens.

In the tradition lens group the performance of the model drops dramatically with fewer than 10 images, indicating that not enough input information is available to rule out all possibilities of the displayed characters. The optical lens group displayed a steady decent of accuracy when fewer images are available, which is expected, as the blur patterns are more complex and the differences between frames more pronounced, so that the images are more precious to the model and contain less overlapping data, and removing any of them will cause losses in accuracy.

D. Adapting Ability

We train the model with fewer groups of data, exposing it to fewer variations of environment parameters, and examine the model's performance in other environments. The results are shown in Figure 9.



(a) Traditional Lens

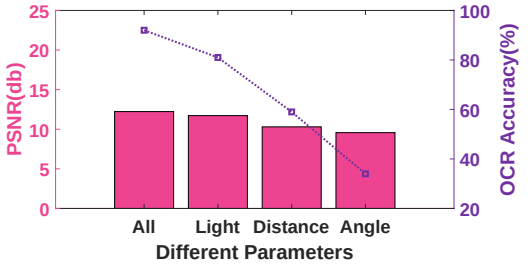


Fig. 9: Performance for Adapting Ability using Traditional Lens and Optical Lens.

We observed that variations in light and angle parameter in training data is crucial to a robust model. In the tradition lens group, variations in distance is not so influential to the results, given that the distances are between 1 and 2 meters. Distance changes do not have such a large impact on the size of the characters shown in the images, so that features extracted from a fixed distance might still exist when these characters vary slightly in size. However, if the network is not exposed to angled images during the training process, the rotations and deformations caused by these angles will easily disturb the feature extraction process. The optical lens group yields similar results except the distance group. As the distance between phones are increased to 6 meters, increases in distance leads to greater complexity of image blurriness and less information, so that the model becomes more sensitive to distance changes.

E. Comparison with other architectures

We train and test other commonly used architectures with the same sets of data and evaluate their results. We chose SRCNN, a commonly used single image SR network, and applied it to each single image before merging the results by pixel-level average. We also used a multi-frame version of CNN consisting of 3D convolutional layers, designed for video super resolution (VideoSR). However, as mentioned above, it is very difficult for the single image approaches to utilize information and distinguish the noisy and deformed patterns, while VideoSR approaches rely upon consistency between frames, so they fail to give satisfactory results. We used the relatively easy 'daytime 1.2m-distance direct with traditional lens' group of data for testing. The results are shown in Table II.

TABLE II: Comparison with existing systems.

System	SRPeek	SRCNN	VideoSR
PSNR	13.32db	7.69db	8.403db
OCR Accuracy	100%	10%	23%

Fig. 10: Illustration of our real-life scenario (Distance between attacker and victim is shortened for demonstration).

We can see that the PSNR of SRPeek is 13.32dB which 73.2% and 58.6% larger than that of SRCNN and VideoSR. For the OCR accuracy, SRPeek can recognize all characters, but SRCNN and VideoSR can only recognize 10% and 23% of them. The results show the efficiency of our SR model in comparison with existing models.

VII. Case Study

A. Accuracy

We build the system on smartphone and evaluate its performance in real-life environments (shown in Fig. 10). We experiment with a Redmi 6A smartphone (with a camera of 13 million pixels) for the attacker and a HUAWEI Mate8 smartphone for the victim. As the telephoto cameras can assist the attacker to see clearly at approximately 3m distance without any aid from SR algorithms, we believe it is insignificant to further extend this distance to judge it as a threat to privacy, so that the following experiments are performed with traditional lens at 1-2m range. We ask 5 human participants to read the reconstructed characters to evaluate the usability of our model. No participants can read the unprocessed images, but all of them can decipher the information on the reconstructed image without much difficulty. The results are shown in Table III.

The results show that Human can read 95%, 85% and 70% contents in home, transport and theater while the OCR accuracy is 100%, 10% and 23%. That verifies human can obtain the most parts of information from the peeking in various environments. In transport, the vibration of smartphone

TABLE III: Recognition accuracy in various scenarios.

Scenarios	Home	Transport	Theater
Baseline	5%	0	0
OCR	100%	10%	23%
Human	95%	85%	70%

TABLE IV: Impact of hand tremors when the camera (attacker phone) and/or target (victim phone) is handheld.

Accuracy	None	Camera	Target	Both
Baseline	5%	0	5%	5%
OCR	95%	85%	80%	80%
Human	95%	85%	80%	85%

and the darker environment can fool the OCR model for content recognition in comparison with human recognition so that leads lower accuracy in transport and theater.

B. Influence of hand tremors

We ask 5 participants to capture images with handheld smartphones, keeping their hand still to their greatest effort (Handheld camera). We process these images and let them read the results. We compare this performance to the data collected on stationary phones to evaluate the influence of hand tremors. We also ask participants to hold a smartphone in their hands and read a piece of text, without other additional instructions (Handheld target). The user may freely interact with the phone when reading. We capture images of the phone at the same time to see how our system deal with a moving target screen. The results are shown in Table IV.

We can see that with the existence of tremors, the recognition accuracy drops from 95% to 85%/80% for both OCR tools and human. We conclude from the results that hand tremors can impact the performance of our system. Although the edges of the phone are notable marks for image alignment, small shifts in the sub-pixel level will cause blurriness in the results of the networks, lowering the readability of the outputs.

C. Success rate in different tasks

We test the success rate of obtaining crucial information when the observed participant performs several tasks on a phone: reading text message, typing text message, entering PIN, and typing password with numbers, English and special characters. The observed participant will turn off the screen of his phone as soon as he/she finishes the task, while the observing participant will observe through our APP, constantly capturing images and processing them to display a real-time and magnified view of the observed phone. For the first two tasks, we ask the observing participant several questions to test if he/she have collected the vital information (e.g. the name or the location mentioned in the text). For the task of recognizing PIN and passwords, we use accuracy per character as a supplementary evaluation metric. In these two scenarios, we ignore the virtual keyboard and view only the textbox, assuming that the last character of the entered string is always visible. Fewer photos will be available for each character but deciphering English characters is also easier than Chinese characters, and we trained a model specifically for each task (with the same neural network architecture), the training images containing only English characters and numbers (or only numbers for the PIN entry scenario). The results are shown in Table V.

TABLE V: Success rate in different tasks.

Accuracy	Read text	Type text	Enter PIN	Enter password
Raw Image	5%	0	0	0
SRPeek	100%	100%	100%	80%

We prove from this experiment that our system functions normally in everyday scenarios and poses a threat to screen privacy.

D. Perceived shoulder surfing susceptibility

We ask the observed participant to rate the perceived shoulder-surfing susceptibility in these scenarios. The attacker will be sitting or standing behind the participant at 1.8m range, pretending to be interacting with their own phone while continuously running the shoulder-surfing APP. None of the participants were alerted by the attacker's behavior and reported suspicion of shoulder-surfing. We believe that our system can enable a malicious attacker to gather large amounts of critical information from the victim while remaining unnoticed.

VIII. Limitations & Discussions

There are certain limitations of this work. We require a certain degree of image capturing and processing abilities of the attacker's phone, and we also expect the victim not to exert too much disturbance to the target phone.

Image Capturing Ability Latest models of smartphones can capture images at 10 frames per second in burst mode easily, however, this ability is not common in phones that are 3 years old. Heavily used phones also perform less than ideal when capturing images in burst mode. Also, when capturing images, the user needs to hold their phone as still as possible to avoid motion blur and assist image alignment. The user also has to keep a sharp focus on the target phone.

Processing Ability To achieve best performance the user needs a phone with strong processing capabilities to run the neural network at real time. As neural networks have been common place in numerous modern APPs, most phones of the latest generation have upgraded their processing ability to run neural networks, but older versions might not possess such processing powers and cannot process images at real-time.

Motion and Tremors In our experiment we assume the observed user will hold still his/hers phone, and not making interactions too often. However, some users may tilt their phones during interactions—especially when typing with one

hand. Too much tremor will cause motion blur in captured images and misalignment between frames, thus degrading the result.

Although our shoulder surfing system proves to be highly efficient against unprotected screens, there are some simple methods to mitigate this unique threat while not cumbering the user.

Dynamic background. Most multi-frame SR algorithms are designed based on the assumption that all input images are reflections of the same scene, and ours is not an exception. By deploying a dynamic background behind the characters, such as tiny dots and lines travelling slowly around the screen, we can construct a constantly changing scene that will confuse the multi-frame SR algorithms, and due to the blurriness of the images, these influential elements cannot be easily removed. These dots do not need to be distinct or colored same as the texts, as multi-frame SR algorithms function with tiny, pixel level differences between frames, making them especially sensitive to microscopic changes.

Active scanning. There are several works providing an active countermeasure against shoulder surfing threats with the naked eye. With front-facing cameras and face detection algorithms, the smartphone can constantly scan the surrounding passers-by and detect their gaze direction, and give a warning to the user when that gaze points at the screen. However, to the extent of our knowledge none of these works have included cameras into their detection scope, but we believe it's practical to implement such features.

Adversarial machine learning methods. In recent years we have discovered the weaknesses of neural networks and that inserting certain microscopic changes, undetectable to the human eye, to the pixels of an image will make it look different to a neural network. These methods fool the feature extraction phases of neural networks, so that SRPeek is also vulnerable to this attack. Theoretically, by exerting a pattern to the victim's screen, it can be captured by the attacker's camera and confuse its SR algorithms, but these remains to be implemented in our future work.

IX. Conclusion

In this work we designed a holistic system, SRPeek, for shoulder surfing on smartphone, which serves as an up-to-date version of a threat model for shoulder surfing, and proved its efficiency. We proved that this threat towards screen privacy is imminent and can steal critical information, including personal texts or passwords, from long distances, thus escaping detection. It is our wish that this work can stir discussion in the field of screen privacy protection and propagate defense mechanisms across critical mobile apps.

The core of SRPeek is a specially designed multi-frame SR network. With its innovative architecture this network outperforms other algorithms of the same field in our application. The design ideology enables this network to process higher levels of data integration ability while keeping a low calculation profile, and we believe the elements of this design can be used in other applications with large amounts of data, such as natural language processing or anomaly detection. Our model can also be used in OCR tasks when multiple images are available, functioning as a preprocessor to improve the quality of the images and increase accuracy.

References

[1] F. Brudy, D. Ledo, and S. Greenberg. Is anyone looking? mediating shoulder surfing on public displays (the video). In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 159–160. 2014.

[2] N. Chakraborty and S. Mondal. Tag digit based honeypot to detect shoulder surfing attack. In *International Symposium on Security in Computing and Communication*, pages 101–110. Springer, 2014.

[3] C.-Y. D. Chen, B.-Y. Lin, J. Wang, and K. G. Shin. Keep others from peeking at your mobile device screen! In *The 25th Annual International Conference*, 2019.

[4] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[5] Z. Dong, S. Zhang, B. Ma, D. Qi, L. Luo, and M. Zhou. A hybrid multi-frame super-resolution algorithm using multi-channel memristive pulse coupled neural network and sparse coding. In *2019 7th International Conference on Information, Communication and Networks (ICICN)*, pages 185–190, 2019.

[6] M. Eiband, M. Khamis, E. Von Zezschwitz, H. Hussmann, and F. Alt. Understanding shoulder surfing in the wild: Stories from users and observers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4254–4265, 2017.

[7] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar. Robust shift and add approach to superresolution. In *Applications of Digital Image Processing XXVI*, volume 5203, pages 121–130. International Society for Optics and Photonics, 2003.

[8] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, page 148–156, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.

[9] W. Goucher. Look behind you: the dangers of shoulder surfing. *Computer Fraud & Security*, 2011(11):17–20, 2011.

[10] Y. Huang, W. Wang, and L. Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):1015–1028, 2017.

[11] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016.

[12] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa, and J. Nalepa. Deep learning for multiple-image super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 2019.

[13] M. Kumar, T. Garfinkel, D. Boneh, and T. Winograd. Reducing shoulder-surfing by using gaze-based password entry. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 13–19, 2007.

[14] T. Kwon, S. Shin, and S. Na. Covert attentional shoulder surfing: Human adversaries are more powerful than expected. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(6):716–727, 2013.

[15] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[16] A. Lucas, S. Lopez-Tapia, R. Molina, and A. K. Katsaggelos. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing*, 28(7):3312–3327, 2019.

[17] J. Lyn and S. Yan. Image super-resolution reconstruction based on attention mechanism and feature fusion. *arXiv preprint arXiv:2004.03939*, 2020.

[18] F. Maggi, A. Volpatto, S. Gasparini, G. Boracchi, and S. Zanero. Poster: Fast, automatic iphone shoulder surfing. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 805–808, 2011.

[19] H. Nasrollahi, K. Farajzadeh, V. Hosseini, E. Zarezadeh, and M. Abdollahzadeh. Deep artifact-free residual network for single-image super-resolution. *Signal, Image and Video Processing*, 14(2):407–415, 2020.

[20] A. Papadopoulos, T. Nguyen, E. Durmus, and N. Memon. Illusionpin: Shoulder-surfing resistant authentication using hybrid images. *IEEE Transactions on Information Forensics and Security*, 12(12):2875–2889, 2017.

[21] Polotiko. Aguirre furious at photo leak of private text message, 2017.

[22] A. Saad, M. Chukwu, and S. Schneegass. Communicating shoulder surfing attacks to users. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, pages 147–152, 2018.

[23] M. S. Sajjadi, R. Vemulapalli, and M. Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6626–6634, 2018.

[24] F. Schaub, R. Deyhle, and M. Weber. Password entry usability and shoulder surfing susceptibility on different smartphone platforms. In *Proceedings of the 11th international conference on mobile and ubiquitous multimedia*, pages 1–10, 2012.

[25] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[26] N. Suetake, M. Sakano, and E. Uchino. Image super-resolution based on local self-similarity. *Optical review*, 15(1):26–30, 2008.

[27] S. Wiedenbeck, J. Waters, L. Sobrado, and J.-C. Birget. Design and evaluation of a shoulder-surfing resistant graphical password scheme. In *Proceedings of the working conference on Advanced visual interfaces*, pages 177–184, 2006.

[28] G. Y. Youm, S. H. Bae, and M. Kim. Image super-resolution based on convolution neural networks using multi-channel input. In *IEEE Image, Video, & Multidimensional Signal Processing Workshop*, 2016.